Machine Learning meets Number Theory: The Data Science of Birch-Swinnerton-Dyer

Laura Alessandretti[1,2], Andrea Baronchelli[2,3], Yang-Hui He[2,4,5]

[1] *Copenhagen Center for Social Data Science, University of Copenhagen, Copenhagen K, 1353, Denmark*
[2] *Department of Mathematics, City, University of London, EC1V 0HB, UK*
[3] *The Alan Turing Institute, London NW1 2DB, UK*
[4] *Merton College, University of Oxford, OX14JD, UK*
[5] *School of Physics, NanKai University, Tianjin, 300071, P.R. China*

l.alessandretti@gmail.com, Andrea.Baronchelli.1@city.ac.uk, hey@maths.ox.ac.uk

Abstract

Empirical analysis is often the first step towards the birth of a conjecture. This is the case of the Birch-Swinnerton-Dyer (BSD) Conjecture describing the rational points on an elliptic curve, one of the most celebrated unsolved problems in mathematics. Here we extend the original empirical approach, to the analysis of the Cremona database of quantities relevant to BSD, inspecting more than 2.5 million elliptic curves by means of the latest techniques in data science, machine-learning and topological data analysis.

Key quantities such as rank, Weierstrass coefficients, period, conductor, Tamagawa number, regulator and order of the Tate-Shafarevich group give rise to a high-dimensional point-cloud whose statistical properties we investigate. We reveal patterns and distributions in the rank versus Weierstrass coefficients, as well as the Beta distribution of the BSD ratio of the quantities. Via gradient boosted trees, machine learning is applied in finding inter-correlation amongst the various quantities. We anticipate that our approach will spark further research on the statistical properties of large datasets in Number Theory and more in general in pure Mathematics.

Contents

1 Introduction and Summary

Elliptic curves $\mathcal{E}$ occupy a central stage in modern mathematics, their geometry and arithmetic providing endless insights into the most profound structures. The celebrated Conjecture of Birch and Swinnerton-Dyer [BSD] is the key result dictating the behaviour of $\mathcal{E}$ over finite number fields and thereby, arithmetic. Despite decades of substantial progress, the proof of the conjecture remains elusive. To gain intuition, a highly explicit and computational programme had been pursued by Cremona [Cre], in cataloguing all elliptic curves up to isogeny and expressed in canonical form, to conductors into the hundreds of thousands.

Interestingly, a somewhat similar situation exists for the higher dimensional analogue of elliptic curves considered as Ricci-flat Kähler manifolds, viz., Calabi-Yau manifolds. Though Yau [Yau] settled the Calabi-Yau Conjecture [Ca], much remains unknown about the landscape of such manifolds, even over the complex numbers. For instance, even seemingly simple questions of whether there is a finite number of topological types of Calabi-Yau n-folds for $n \geq 3$ is not known – even though it is conjectured so. Nevertheless, because of the pivotal importance of Calabi-Yau manifolds to superstring theory, theoretical physicists have been constructing ever-expanding datasets thereof over the last few decades (cf. [HeBook] for a pedagogical

introduction).

Given the recent successes in the science of "big data" and machine learning, it is natural to examine the database of Cremona [Cre2] using the latest techniques of Data Science. Indeed, such a perspective has been undertaken for Calabi-Yau manifolds and the landscape of compactifications in superstring theory in high-energy physics, ranging from machine-learning [He] to statistics [Doug, HJP]. Indeed, [He, KS, CHKN, Rue] brought about a host of new activities in machine-learning within string theory; moreover, [He, HeBook] and the subsequent work in [BHJM, ACHN, HL, JKP, HK, BCDL, **?**] introduced the curious possibility that machine-learning should be applied to at least stochastically avoid expensive algorithms in geometry and combinatorics and to raise new conjectures.

Can artificial intelligence help with understanding the syntax and semantics of mathematics? While such profound questions are better left to the insights of Turing and Voevodsky, the more humble question of using machine-learning to recognizing patterns which might have been missed by standard methods should be addressed more immediately. Preliminary experiments such as being able to "guess"- to over 99% accuracy and confidence - the ranks of cohomology groups without exact-sequence-chasing (having seen tens of thousands of examples of known bundle cohomologies) [HeBook], or whether a finite group is simple without recourse to theorem of Noether and Sylow (having been trained on thousands of Cayley tables) [HK] already point to this potentiality.

In our present case of number theory, extreme care should, of course, be taken. Patterns in the primes can be notoriously deceptive, as exemplified by the likes of Skewe's constant. Indeed, a sanity check to let neural networks predict the next prime number in [He] yielded a reassuring null result. Nevertheless, one should not summarily disregard all experimentation in number theory: after all, the best neural network of the 19th century - the mind of Gauß - was able to pattern-spot $\pi(x)$ to raise the profound Prime Number Theorem years before the discovery of complex analysis to allow its proof.

The purpose of this paper is to open up dialogue between data scientists and number theorists, as the aforementioned works have done for the machine-learning community with geometers and physicists, using Cremona's elliptic curve database as a concrete arena. The organization is as follows. We begin with a rapid introduction

in Section 2, bearing in mind of the diversity of readership, to elliptic curves in light of BSD. In Section 3, we summarize Cremona's database and perform preliminary statistical analyses beyond simple frequency count. Subsequently, Section 4 is devoted to machine-learning various aspects of the BSD quantities and Section 5, to their topological data analyses and persistent homology. Finally, we conclude in Section 5 with the key results and discuss prospects for the future.

2 Elliptic Curves and BSD

Our starting point is the Weierstraß model

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \tag{2.1}$$

of an elliptic curve $\mathcal{E}$ over $\mathbb{Q}$, where $(x, y) \in \mathbb{Q}$ and the coefficients $a_i \in \mathbb{Z}$. The discriminant Δ and J-invariant of $\mathcal{E}$ are obtained in a standard way as

$$\Delta(\mathcal{E}) = -b_2^2b_8 - 8b_4^3 - 27b_6^2 + 9b_2b_4b_6 \ , \qquad j(\mathcal{E}) = \frac{c_4^3}{\Delta} \tag{2.2}$$

where $b_2 := a_1^2 + 4a_2$, $b_4 := 2a_4 + a_1a_3$, $b_6 := a_3^2 + 4a_6$, $b_8 := a_1^2a_6 + 4a_2a_6 - a_1a_3a_4 + a_2a_3^2 - a_4^2$ and $c_4 := b_2^2 - 24b_4$. Smoothness is ensured by the non-vanishing of Δ and isomorphism (isogeny) between two elliptic curves, by the equality of j.

An algorithm of Tate and Laska [Tat, Las] * can then be used to bring the first 3 coefficients $a_{1,2,3}$ to be $0, \pm 1$, transforming (2.1) into a *minimal Weierstraß model*.

* In particular, consider the transformation between the coefficiens a_i and a'_i between two elliptic curves $\mathcal{E}$ and $\mathcal{E}'$:

$$\begin{aligned}
ua'_1 &= a_1 + 2s,\\
u^2a'_2 &= a_2 - sa_1 + 3r - s^2,\\
u^3a'_3 &= a_3 + ra_1 + 2t,\\
u^4a'_4 &= a_4 - sa_3 + 2ra_2 - (t + rs)a_1 + 3r^2 - 2st,\\
u^6a'_6 &= a_6 + ra_4 + r^2a_2 + r^3 - ta_3 - t^2 - rta_1.
\end{aligned}$$

for $u, s, t \in \mathbb{Q}$, then relating the points $(x, y) \in \mathcal{E}$ and $(x', y') \in \mathcal{E}'$ as

$$x = u^2x' + r \ , \quad y = u^3y' + su^2x' + t$$

yields $u^{12}\Delta' = \Delta$ and hence $j' = j$, and thus the isomorphism.

Thus, for our purposes, an elliptic curve $\mathcal{E}$ is specified by the pair of integers (a_4, a_6) together with a triple $(a_1, a_2, a_3) \in \{-1, 0, 1\}$. From the vast subject of elliptic curves, we will need this 5-tuple, together with arithmetic data to be presented in the ensuing.

2.1 Rudiments on the Arithmetic of $\mathcal{E}$

This subsection serves as essentially a lexicon for the quantities which we require, presented in brief, itemized form.

Conductor and Good/Bad Reduction: The conductor is the product over all (finite many) number of primes p - the primes of bad reduction - where $\mathcal{E}$ reduced modulo p becomes singular (where $\Delta = 0$). All other primes are called good reduction.

Rank and Torsion: The set of rational points on $\mathcal{E}$ has the structure of an Abelian group, $\mathcal{E}(\mathbb{Q}) \simeq \mathbb{Z}^r \times T$. The non-negative integer r is called the rank, its non-vanishing would signify an infinite number of rational points on $\mathcal{E}$. The group T is called the torsion group and can be only one of 15 groups by Mazur's celebrated theorem [Maz], viz., the cyclic group C_n for $1 \leq n \leq 10$ and $n = 12$, as well as the direct product $C_2 \times C_n$ for $n = 2, 4, 6, 8$.

L-Function and Conductor: The Hasse-Weil Zeta-function of $\mathcal{E}$ can be defined, given a finite field $\mathbb{F}_{q=p^n}$, as the generating functions z_p (the local) and Z (the global):

$$\begin{aligned} Z_p(t; \mathcal{E}) &:= \exp\left(\sum_{n=1}^{\infty} \frac{\mathcal{E}(\mathbb{F}_{p^n})}{n} t^n\right) , \\ Z(s; \mathcal{E}) &:= \prod_p Z_p(t := p^{-s}; \mathcal{E}) \ . \end{aligned} \tag{2.3}$$

Here, in the local zeta function $Z_p(t; \mathcal{E})$, $\mathcal{E}(\mathbb{F}_{p^n})$ is the number of points of $\mathcal{E}$ over the finite field and the product is taken over all primes p to give the global zeta function $Z(s)$.

The definition (2.3) is applicable to general varieties, and for elliptic curves, the global zeta function simplifies (cf. [Sil]) to a product of the Riemann zeta

function ζ and a so-called L-function as

$$Z(s;\mathcal{E}) = \frac{\zeta(s)\zeta(s-1)}{L(s;\mathcal{E})} \ , \tag{2.4}$$

where

$$L(s;\mathcal{E}) = \prod_p L_p(s;\mathcal{E})^{-1} \ , \quad L_p(s;\mathcal{E}) := \begin{cases} (1-\alpha_p p^{-s}+p^{1-2s}), & p \nmid N \\ (1-\alpha_p p^{-s}), & p \mid N \text{ and } p^2 \nmid N \\ 1, & p^2 \mid N \end{cases} \ .$$

In the above, $\alpha_p = p + 1 -$ counts the number of points of $\mathcal{E}$ mod p for primes of good reduction and ± 1 depending on type of bad reduction. The positive integer N which controls, via its factorization, these primes, is the conductor of $\mathcal{E}$.

Importantly, the L-function has analytic continuation [TW] to $\mathbb{C}$ so that the variable s is not a merely dummy variable like t in the local generating function, but renders $L(s;\mathcal{E})$ a complex analytic function.

Real Period: The periods of a complex variety is usually defined to be the integral of some globally defined holomorphic differential over a basis in homology. Here, we are interested in the real period, defined as (using the minimal Weierstraß model)

$$\mathbb{R} \ni \Omega := \int_{\mathcal{E}(\mathbb{R})} |\omega| \ , \quad \omega = \frac{dx}{2y + a_1 x + a_3} \tag{2.5}$$

over the set of real points $\mathcal{E}(\mathbb{R})$ of the elliptic curve.

Tamagawa Number: Now, $\mathcal{E}$ over any field is a group, thus in particular we can define $\mathcal{E}(\mathbb{Q}_p)$ over the p-adic field $\mathbb{Q}_p$ for a given prime, as well as its subgroup $\mathcal{E}^0(\mathbb{Q}_p)$ of points which have good reduction. We define the index in the sense of groups

$$c_p := \Big[\mathcal{E}(\mathbb{Q}_p) : \mathcal{E}(\mathbb{Q}_p)\Big] \ , \tag{2.6}$$

which is clearly equal to 1 for primes of good reduction since then $\mathcal{E}^0(\mathbb{Q}_p) = \mathcal{E}(\mathbb{Q}_p)$. The Tamagawa number is defined to be the product over all primes of bad-reduction of c_p, i.e.,

$$\text{Tamagawa Number } = \prod_{p|N} c_p \ . \tag{2.7}$$

Canonical Height: For a rational point $P = \frac{a}{b}$, written in minimal fraction form with $\gcd(a,b) = 1$, $a, b \in \mathbb{Z}$ and $b > 0$, we can define a naive height $h(P) := \log \max(|a|, b)$. Then, a *canonical height* can be defined as

$$\hat{h}(P) = \lim_{n\to\infty} n^{-2} h(nP) \ , \tag{2.8}$$

where $nP = P + \ldots + P$ (n-times) is the addition of under the group law of $\mathcal{E}$. This limit exists and renders $\hat{h}$ the unique quadratic form on $\mathcal{E}(\mathbb{Q}) \otimes \mathbb{R}$ such that $\hat{h} - h$ is bounded. An explicit expression of $\hat{h}(P)$ in terms of a, b can be found, e.g., in Eq4 and Eq5 of [BGZ]. The canonical height defines a natural bilinear form

$$2\langle P, P' \rangle = \hat{h}(P + P') - \hat{h}(P) - \hat{h}(P') \tag{2.9}$$

for two points $P, P' \in \mathcal{E}(\mathbb{Q})$ and as always, $P + P'$ is done via the group law.

Regulartor: Given the infinite (free Abelian) part of $\mathcal{E}(\mathbb{Q})$, viz., $\mathbb{Z}^r$, let its generators be $P_1, \ldots, P_r$, then we can define the regulator

$$R_{\mathcal{E}} = \det \langle P_i, P_j \rangle \ , \quad i, j = 1, \ldots, r \tag{2.10}$$

where the pairing is with the canonical height and defines an $r \times r$ integer matrix. For $r = 0$, R is taken to be 1 by convention.

Tate-Shafarevich Group: Finally, one defines group cohomologies $H^1(\mathbb{Q}, \mathcal{E})$ and $H^1(\mathbb{Q}_p, \mathcal{E})$ between which there is a homomorphism (cf. e.g., [Mil]IV.2 for a detailed description). We can then define the Tate-Shafarevich Group Ш of $\mathcal{E}$ as the kernel of the homomorphism

$$Ш(\mathcal{E}) := \ker \left(H^1(\mathbb{Q}, \mathcal{E}) \longrightarrow \prod_p H^1(\mathbb{Q}_p, \mathcal{E}) \right) \ . \tag{2.11}$$

This is the most mysterious part of the arithmetic of elliptic curves, it is conjectured to be a finite Abelian group. For ranks $r = 0, 1$, this has been proven (cf. the survey of [RS]) but in general this is not known.

2.2 The Conjecture

With the above definitions, we can at least present the celebrated

CONJECTURE 1 (Birch–Swinnerton-Dyer (Weak Version)) *The order of the zero of $L(s;\mathcal{E})$ at $s=1$ is equal to the rank r,*

$$\operatorname*{Ord}_{s\to 1} L(s;\mathcal{E}) = r(\mathcal{E}) \ .$$

That is, the Taylor series around 1 is $L(s;\mathcal{E}) \sim c(s-1)^r$ with some complex coefficient c.

In fact, a stronger version of the conjecture predicts precisely what the Taylor coefficient c should be:

CONJECTURE 2 (Birch–Swinnerton-Dyer (Strong Version)) *The Taylor coefficient of $L(s;\mathcal{E})$ at $s=1$ is given in terms of the regulator R, Tamagawa number $\prod_{p|N} c_p$, (analytic) order of the Tate-Shafarevich group $Ш$, and the order of the torsion group T. Specifically, let $L(s;\mathcal{E}) = \sum_r \frac{L^{(r)}(1;\mathcal{E})}{r!}(s-1)^r$, then*

$$\frac{L^{(r)}(1;\mathcal{E})}{r!} = \frac{|Ш|\cdot\Omega\cdot R\cdot \prod_{p|N} c_p}{|T|^2} \ .$$

3 Elliptic Curve Data

BSD arose from extensive computer experimentation, the earliest of its kind, by Birch and Swinnerton-Dyer. Continuing along this vein, Cremona [Cre] then compiled an impressive of list of 2,483,649 isomorphism classes of elliptic curves over $\mathbb{Q}$ and explicitly computed the relevant quantities introduced above. This is available freely online at [Cre2].

3.1 Cremona Database

The database of Cremona, on which the rest of the paper will focus, associates to each minimal Weierstraß model (given by the coefficients $(a_1,a_2,a_3) \in \{-1,0,1\}$ and $(a_4,a_6) \in \mathbb{Z}$; generically, these last two coefficients have very large magnitude), the following:

- the conductor N, ranging from 1 to 400,000;
- the rank r, ranging from 0 to 4;
- the torsion group T, whose size ranges from 1 to 16;
- the real period Ω, a real number ranging from approximately $2.5 \cdot 10^{-4}$ to 6.53.
- the Tamagawa number $\prod_{p|N} c_p$, ranging from 1 to 87040;
- the order of the Tate-Shafarevich group (exactly when known, otherwise given numerically), ranging from 1 to 2500;
- the regulator $R \in \mathbb{Z}_{>0}$, ranging from approximately 0.01 to 3905.84.

A typical entry, corresponding to the curve $y^2+xy = x^3-x^2-453981x+117847851$ (labelled as "314226b1" and with $\Delta = 2 \cdot 3^3 \cdot 11 \cdot 23^8$ and $j = 2^{-1} \cdot 3^3 \cdot 11^{-1} \cdot 23 \cdot 199^3$ which are readily computed from(2.2)) would be

$$(a_1, a_2, a_3, a_4, a_6) = (1, -1, 0, -453981, 117847851) \qquad \Longrightarrow$$

$$\begin{cases} N &= 314226 = 2 \cdot 3^3 \cdot 11 \cdot 23^2 \\ r &= 0 \\ R &= 1 \\ \Omega &\simeq 0.56262 \\ \prod_{p|N} c_p &= 3 \\ |T| &= 3 \\ |Ш| &= 1 \ . \end{cases} \tag{3.1}$$

3.2 Weierstraß Coefficients

Let us begin with a statistical analysis of the minimal Weierstraß coefficients themselves. It so happens that in the entire database, there are only 12 different sets of values of (a_1, a_2, a_3), we tally all of the curves in the following histogram, against rank

and (a_1, a_2, a_3):

$(a_1, a_2, a_3)\backslash$Rank	0	1	2	3	4
$\{0, -1, 0\}$	126135	155604	30236	659	0
$\{0, -1, 1\}$	17238	24593	7582	399	0
$\{0, 0, 0\}$	172238	213780	40731	698	0
$\{0, 0, 1\}$	28440	39235	11187	506	0
$\{0, 1, 0\}$	118942	157003	34585	722	0
$\{0, 1, 1\}$	18016	27360	9609	426	0
$\{1, -1, 0\}$	102769	127198	25793	551	1
$\{1, -1, 1\}$	96995	128957	28940	604	0
$\{1, 0, 0\}$	66411	98092	25286	612	0
$\{1, 0, 1\}$	71309	94595	20907	548	0
$\{1, 1, 0\}$	69759	88403	18293	496	0
$\{1, 1, 1\}$	67834	91717	21197	458	0

We see that most of the curves are of smaller rank, with only a single instance of $r = 4$. This is in line with the recent result of [BS] that most elliptic curves are rank 1; in fact, over 2/3 of elliptic curves obey the BSD conjecture [BSZ].

To give an idea of the size of the a-coefficients involved, the largest one involved in the database is

$$\vec{a} = \{1, 0, 0, -40101356069987968, -309091244068737325444480 0\} \ , \tag{3.2}$$

which is of rank 0.

Even though it is a conjecture that the rank r can be arbitrarily large, the largest confirmed rank [Elk] so far known in the literature is 19, corresponding (the last term being a 72-digit integer!) to

$$a_1 = 1,\ a_2 = 1,\ a_3 = -1,\ a_4 = 31368015812338065133318565292206590792820353345,$$
$$a_6 = 302038802698566087335643188429543498624522041683874493555186062568159847 \ . \tag{3.3}$$

This extraordinary result is clearly not in the database due to the large rank.

One of the first steps in data visualization is a *principle component analysis* where features of the largest variations are extracted. The minimal Weierstraß model gives a natural way of doing so, since the (a_1, a_2, a_3) coefficients take only 12 values and we can readily see scatter plot of (a_4, a_6). Now, due to the large variation in these coefficients, we define a signed natural logarithm for $x \in \mathbb{R}$ as

$$\mathrm{sLog}(x) = \begin{cases} \mathrm{sgn}(x)\log(x)\ , & x \neq 0 \\ 0\ , & x = 0\ . \end{cases} \tag{3.4}$$

We present this scatter plot of $(\mathrm{sLog}(a_4),\ \mathrm{sLog}(a_6))$ in Fig. 1. Therein, we plot all the data points (i.e., for all different values of (a_1, a_2, a_3)) together, distinguishing rank by colour (rank 4 has only a single point as seen from the table above).

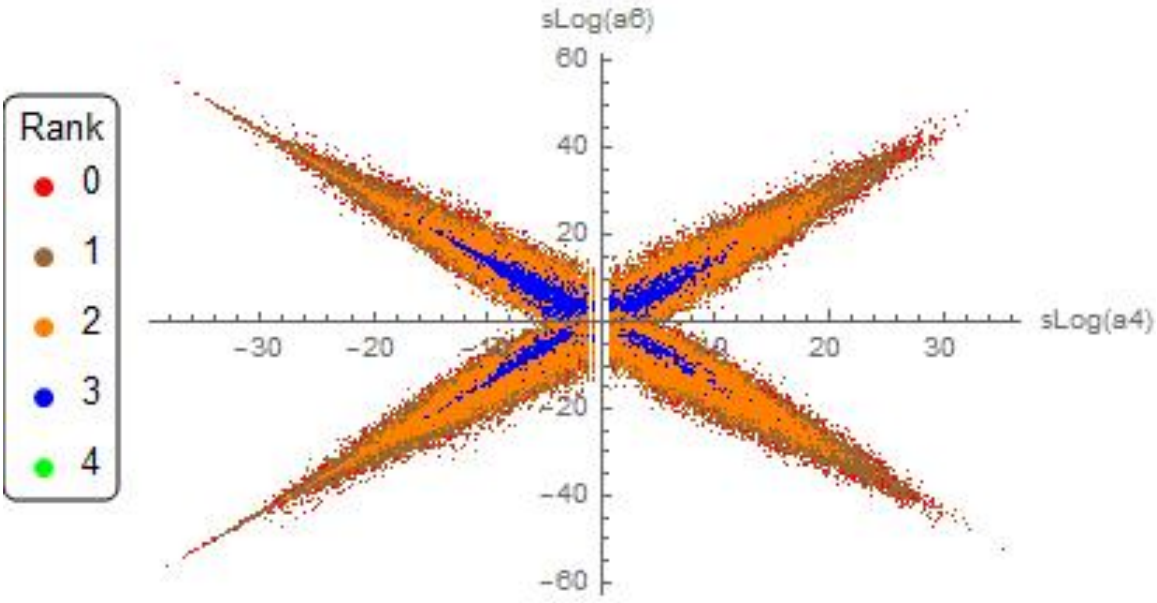


Figure 1: A scatter plot of $(\mathrm{sLog}(a_4),\ \mathrm{sLog}(a_6))$ for all 2,483,649 elliptic curves in the Cremona database. Different ranks are marked with different colours.

The first thing one would notice is the approximate cross-like symmetry, even within rank. However, this is not trivial because the transformation

$$(a_4, a_6) \longrightarrow (\pm a_4, \pm a_6) \tag{3.5}$$

is by no means a rank preserving map. For instance, a single change in sign in a_4,

could result in rank change from 1 to 3:

$$r(\{0,1,1,-10,20\}) = 3\ , \quad r(\{0,1,1,+10,20\}) = 1\ . \tag{3.6}$$

Examples of a similar nature abound. The next feature to notice is that the size of the cross shrinks as the rank increases. This is rather curious since the largest rank case of (3.3) has far *larger* coefficients. This symmetry is somewhat reminiscent of mirror symmetry for Calabi-Yau 3-folds, where every compact smooth such manifold with Hodge numbers $(h^{1,1}, h^{2,1})$ has a mirror manifold with these values exchanged.

3.3 Distributions of a_4 and a_6

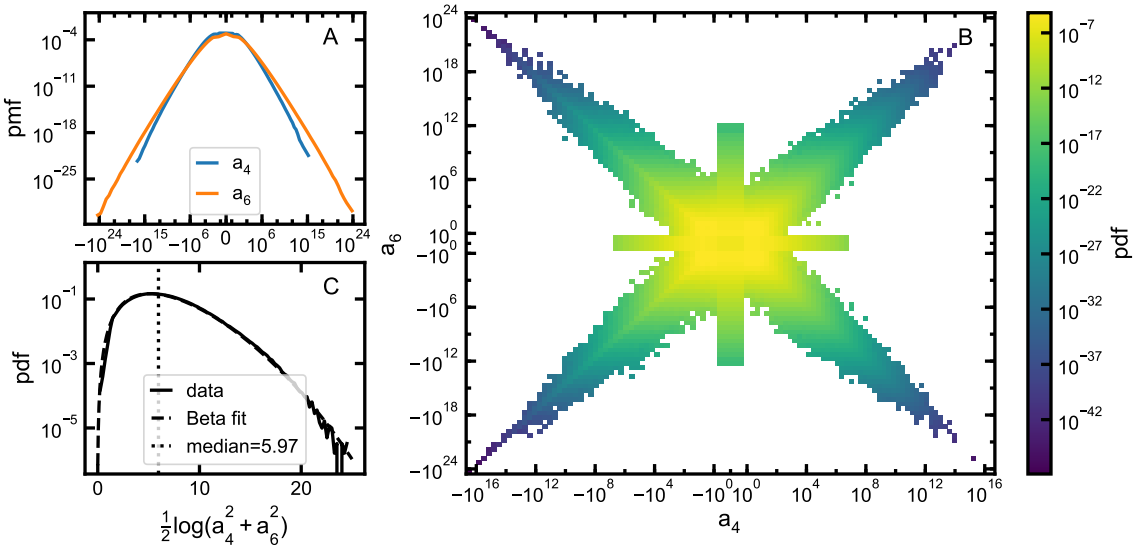


Figure 2: **The distributions of a_4 and a_6.** (A) Probability mass of a_4 (blue line) and a_6 (orange line). (B) Joint probability of a_4 and a_6. The colour indicates the density of points within each bin (see colour-bar). Note that the figure axes are in symlog scale (linear scale between -1 and 1, logarithmic scale for other values in the range). Hence, we can see also the density corresponding to $a_4 = 0$ and $a_6 = 0$ (cross in the middle). (C) Probability density distribution for the logarithm(in base 10) of $\sqrt{a_4^2 + a_6^2}$ (when $a_4 > 0$ and $a_6 > 0$, filled line), the corresponding Beta distribution fit with parameters $\alpha = 4.1$, $\beta = 25.0$ and $s = 44.1$ (dashed line), and the median value (dotted line).

Fortified by the above initial observations, let us continued with more refined study of the distribution of a_4 and a_6. First, let us plot the distribution of each individually, normalized by the total, i.e., as *probability mass functions*. These are

shown in part (A) to the left of Fig 2. Note that the horizontal axes (binning) is done logarithmically. We see that the distributions of both a_4 and a_6 are symmetric with respect to 0 (Fig. 2-A), with a_4 spanning ~ 8 orders of magnitude smaller as compared to a_6. This is just to give an idea of the balanced nature of Cremona's data, that elliptic curves with $\pm a_4$ and $\pm a_6$ are all constructed.

Next, in part (B) of Fig 2, we plot the joint probability mass function of the pair (a_4, a_6) with colour showing the frequency as indicated by the colour-bar to the right. We see that, as discussed in Fig. 1, there is a cross-like symmetry. Here, since we are not separating by rank, the symmetry is merely a reflection of the constructions of the dataset, that $\pm a_4$ and $\pm a_6$ are all present. What is less explicable is that it should be a cross shape and what is the meaning of the boundary curve beyond which there does not seem to be any minimal models. For reference, the central rectilinear cross indicates the cases of $a_4 = 0$ and $a_6 = 0$ respectively.

Finally, we compute the Euclidean distance $d := \sqrt{a_4^2 + a_6^2}$ from the origin and study its probability distribution. This is shown in part (C) of Fig. 2, We find that half of the data lies within a radius of $\sim 10^6$ from the origin. The logarithm of d can be well fitted with a Beta probability distribution:

$$f(x, \alpha, \beta, s) = K \cdot \frac{x}{s}^{\alpha-1} \left(1 - \frac{x}{s}\right)^{\beta-1} , \tag{3.7}$$

with parameters $\alpha = 4.1$, $\beta = 25.0$ and $s = 44.1$. Thus, whilst there are a number of coefficients of enormous magnitude, the majority still have relatively small ones.

Differences by Rank: As with our initial observations, we now study the variation of (a_4, a_6) with respect to the rank r of the elliptic curves. First, in Fig. 3 parts A-D, we plot the joint distributions of a_4 and a_6 for $r = 0, 1, 2, 3$ respectively. We can see that they differ signigicantly from each other, under permutation test [PJ] at confidence level $\alpha = 0.01$.

Next, Fig. 3 E show the probability distribution functions for our Euclidean distance $\sqrt{a_4^2 + a_6^2}$ for the different ranks. We find that the median Euclidean distance from the center decreases for higher values of r. In fact, we see that the median values of a_4 and a_6 increase with the rank r (see Fig. 4D-E which we will discuss shortly). Again, each is individually well-fitted by the Gamma distribution. In tables 9 and 10

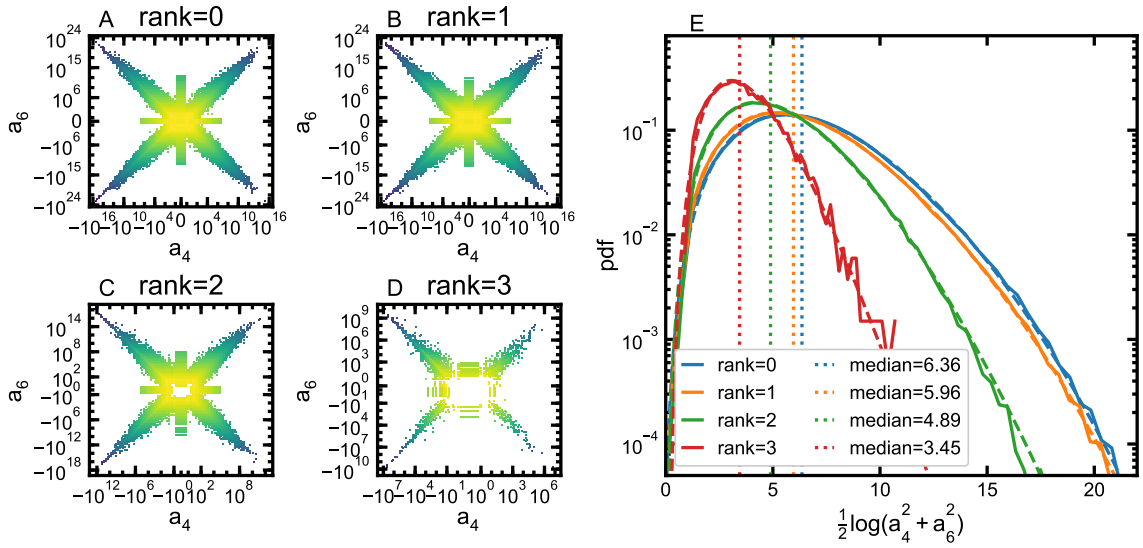


Figure 3: **Different distributions of a_4 and a_6 for different ranks.** (A-D) Joint probability distribution of a_4 and a_6 for values of the rank $r = 0$ (A), $r = 1$ (B), $r = 2$ (C), $r = 3$ (D). (E) Probability density distribution for the logarithm(in base 10) of $\sqrt{a_4^2 + a_6^2}$ (when $a_4 > 0$ and $a_6 > 0$, filled lines) for various value of the rank r, the corresponding Beta distribution fit (dashed lines), and the corresponding median values (dotted lines).

in the Appendix, we show some statistics of a_4 and a_6 including their mean, standard deviation, median, and the number of zero entries, for given rank r and values of (a_1, a_2, a_3).

3.4 Distributions of various BSD Quantities

Now, the coefficients a_i are the inputs of the dataset, each specifying a minimal model of an elliptic curve, and to each should be associated the numerical tuple $(r, N, R, \Omega, \prod_{p|N} c_p, |T|, |Ш|)$ for the rank, the conductor, the regulator, the real period, the Tamagawa number, the order of the torsion group and the order of Tate-Shafarevich group, respectively. It is now expedient to examine the distribution of "output" parametres.

As always, we arrange everything by rank $r = 0, 1, 2, 3$ and in Fig. 4 show the box plots of the variation around the median (drawn in red). The boxes enclose 50% of the distribution, and the whiskers, 95% of the distribution. We see that, as detailed

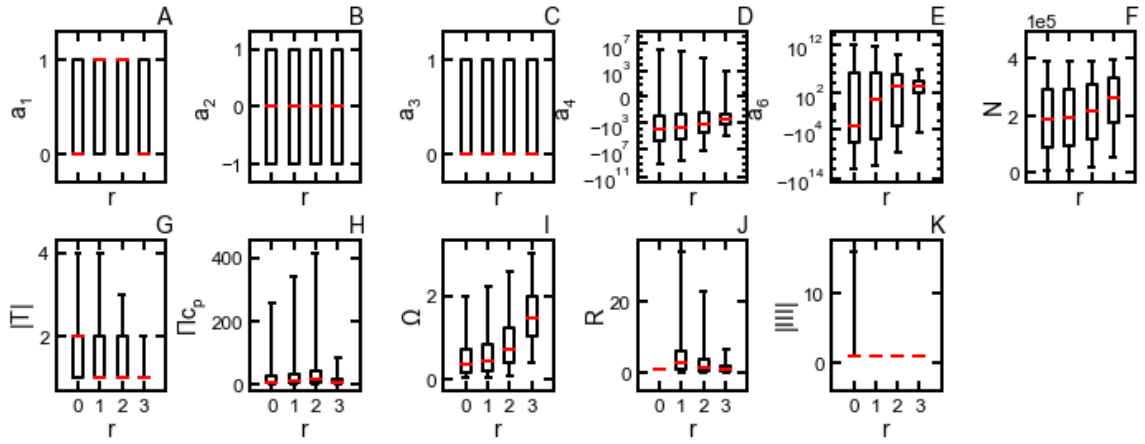


Figure 4: **Characteristics of the elliptic curves based on their rank.** Boxplots of a_1, a_2, a_3, a_4, a_6 in parts (A)-(E) respectively; boxplots of N, $|T|$, $\prod_{p|N} c_p$, Ω, R and $|Ш|$ in parts (F)-(K) respectively, all for different values of the rank $r = 0, 1, 2, 3$. The red line shows the median value, the boxes enclose 50% of the distribution, and the whiskers, 95% of the distribution.

above, $a_{1,2,3}$ have only variation $[-1, 1]$ and a_4 has many orders of magnitude more in variation that a_6. The conductor N has a fairly tame distribution while the other BSD quantities vary rather wildly – the is part of the difficulty of the conjecture, the relevant quantities behave quite unpredictably.

The RHS of Conjecture 2: We now put all the quantities together according to the RHS of the Strong BSD Conjecture, which we recall to be $RHS = \frac{(\Omega \cdot R \cdot \prod_{p|N} c_p \cdot |Ш|)}{T^2}$. We test which statistical distribution best describe this data, by comparing 85 continuous distribution under the Akaike information criterion. We find that the distribution best describing the data is the Beta distribution (see Figure 5 A):

$$f(x, a, b) = \frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)} \tag{3.8}$$

where Γ is the standard gamma function, $a = 1.55$, $b = 14.28$, and the original variable has been re-scaled such that $x = RHS/62.71$.

The selected distribution changes for elliptic with a specific rank r. For $r = 0$, the selected distribution is the exponentiated Weibull, while for larger values of r, the selected distribution is the Johnson SB (see Figure 5 B). We find that the median value of the RHS increases both as a function of the rank r (see Figure 5 C) and N

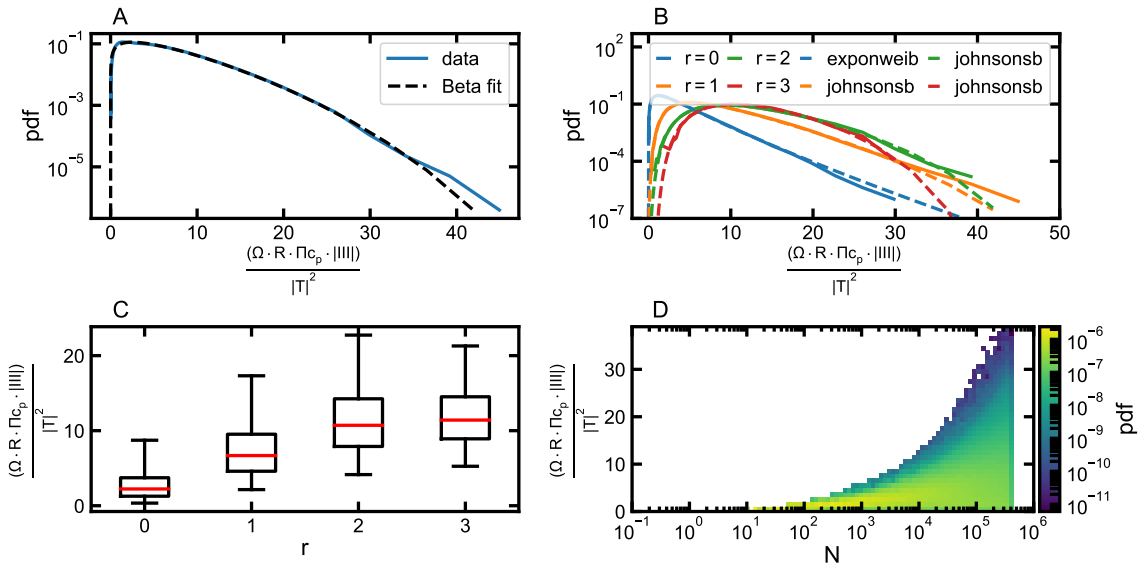


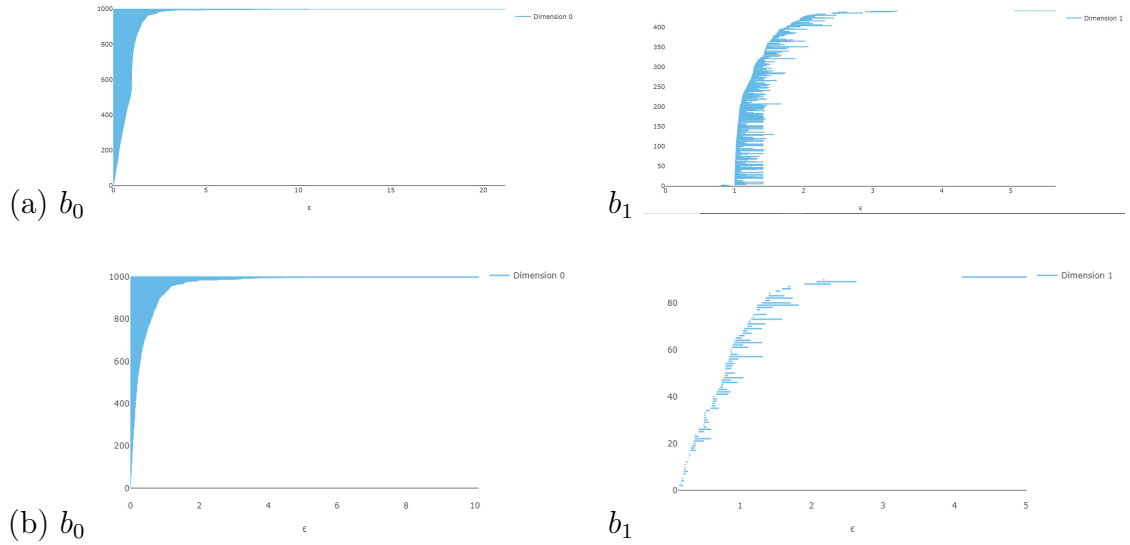
Figure 5: **The RHS of Conjecture 2.** (A) Probability density function of the RHS (filled blue line) and the corresponding Beta distribution fit (dashed black line). (B) Probability density function of the RHS for different values of the rank r (filled lines), and the corresponding best fits chosen with Akaike information criterion (dashed lines) (C) Boxplot showing the value of RHS for different values of r. The red line shows the median value, the boxes to the 50% of the distribution, and the whiskers to the 95% of the distribution.

(see Figure 5 D).

4 Topological Data Analysis

Let us gain some further intuition by visualizing the data. As far as the data is concerned, to each elliptic curve, specified by the Weierstraß coefficients, one associates a list of quantities, the conductor, the rank, the real-period, etc. The define a point cloud in Euclidean space of rather high dimension, each of point of which is defined by the list of these quantities which enter BSD. In the above, we have extracted the last two coefficients of the Weierstraß form of the elliptic curves and studied them against the variations of the first three which, in normal form, can only be one of the 9 possible 3-tuples of ± 1, and 0. The normal form thus conveniently allows us to at least "see" the Weierstraß coefficients because we have projected to two dimensions. However, the full list of the relevant quantities for the elliptic curve has quite a number of entries and cannot be visualized directly.

Luckily, there is precisely a recently developed method in data science which allows for the visualization of "high dimensionality", viz., persistent homolgy in topological data analysis [CZCG] (cf. an excellent introductory survey of [OPTGH]). In brief, one creates a Vietoris-Rips simplex from the data points in Euclidean space, with a notion of neighbourhood ϵ (by Euclidean distance). The Betti numbers b_i of the simplex is then computed as one varies ϵ, whether the values are non-zero for each i gives a notion of whether non-trivial topology (such as holes) persists for different scales ϵ. The result is a so-called **barcode** for the data. In the ensuing, we will use G. Henselman's nice implementation of the standard methods in topological data analysis, the package Eirene for Julia/Python [Ei].



(a) b_0 b_1

(b) b_0 b_1

Figure 6: The barcodes for (a) all the Weierstraß coefficients at Betti numbers 0 and 1; (b) on the (principle component) coefficients $(\mathrm{sLog}(a_4),\ \mathrm{sLog}(a_6))$.

4.1 The Weierstraß Coefficients

We begin by studying the barcodes for the full set of five Weierstraß coefficients a_i (as always, we take sLog for a_4 and a_6 due to their size). Of course, computing the homology of the Vietoris-Rips complex for over 2 million points is computationally impossible. The standard method is to consider random samples. Moreover, the first two Betti numbers b_0 and b_1 are usually sufficiently illustrative. Thus, we will take 1000 random samples (we do not separate the ranks since we find there is no significant difference for the barcodes amongst different ranks) of the coefficients (with the usual sLog for the last two). The barcodes are shown in part (a) of Figure 6. For

reference, we also plot the barcodes for the pair $(\mathrm{sLog}(a_4),\ \mathrm{sLog}(a_6))$ only since $a_{1,2,3}$ do not vary so much.

4.2 A 6-dimensional Point-Cloud

Let us now try to visualize the relevant BSD quantities $(N,\ r,\ |T|,\ \prod_{p|N} c_p,\ \Omega,\ R, Ш)$ together. Organized by the ranks $r = 0, 1, 2$ which dominate the data by far, the 6-tuple

$$(N,\ |T|,\ \prod_{p|N} c_p,\ \Omega,\ R, Ш), \qquad r = 0, 1, 2 \tag{4.9}$$

naturally form three point-clouds in $\mathbb{R}^6$. Due to the high dimensionality, we sample 100 random points for each of the r values and compute the full barcodes $b_{0,\ldots,6}$. It turns out that the main visible features are in dimension 0. We present these in Figure 7 and observe that indeed there is some variation in the barcode amongst the different ranks.

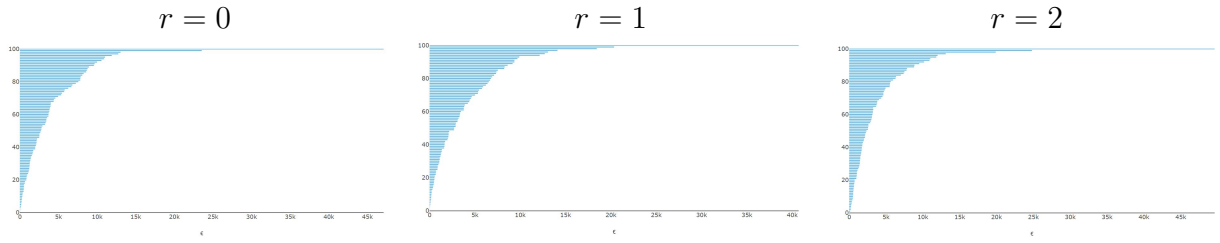


Figure 7: The barcodes for 6-dimensional point-cloud $(N,\ |T|,\ \prod_{p|N} c_p,\ \Omega,\ R, Ш)$, with 100 random samples, for $r = 0, 1, 2$.

4.3 Conductor Divisibility

The factors of the conductor N is of defining importance in the L-function, which appear to the LHS of BSD, meanwhile, the RHS is governed, in the strong case, by the combination $F := \frac{|Ш| \cdot \Omega \cdot R \cdot \prod_{p|N} c_p}{|T|^2}$. It is therefore expedient to consider the point cloud of the 3-tuple (N, r, F) organized by divisibility properties of N. For instance, one could contrast the barcodes for the triple for N even versus N odd. Again, the features are prominent for dimension 0 and the barcodes are shown in Figure 8.

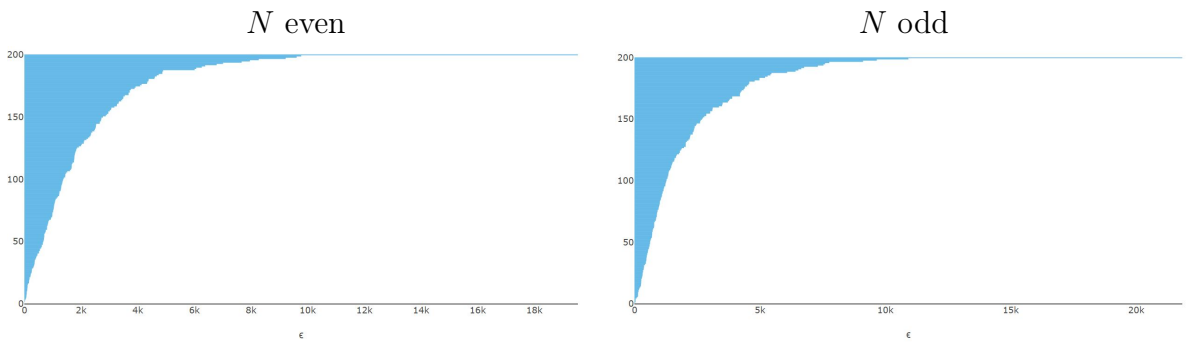


Figure 8: The barcodes for 3-dimensional point-cloud $(N,\ r,\ \frac{|Ш|\cdot\Omega\cdot R\cdot\prod_{p|N} c_p}{|T|^2})$, with 200 random samples for even/odd N.

Simiarly, we could group by N modulo 3, as shown in Figure 9.

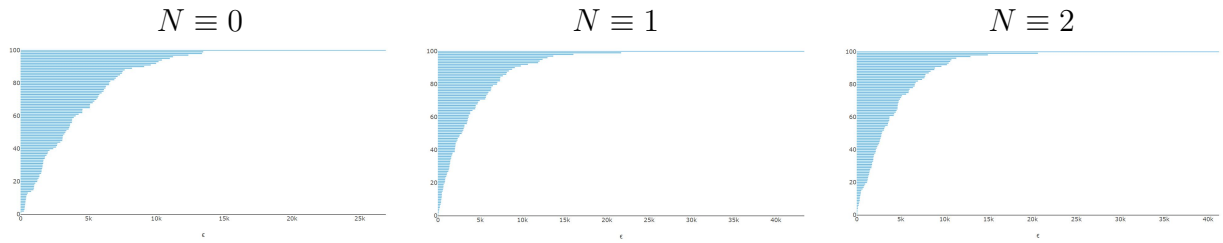


Figure 9: The barcodes at dimension 0 for 3-dimensional point-cloud $(N,\ r,\ \frac{|Ш|\cdot\Omega\cdot R\cdot\prod_{p|N} c_p}{|T|^2})$, with 100 random samples, for N distinguished modulo 3.

5 Machine Learning

In [He, HeBook], a paradigm was proposed to using machine-learning, and in particular deep neural networks to help computations in various problems in algebraic geometry. Exemplified by computing cohomology of vector bundles, it was shown that to very high precision, AI can guess the correct answer without using the standard method of Gröbner basis construction and chasing long exact sequences, both of which are computationally intensive. Likewise, [HK] showed that machine-learning can identify algebraic structures such as distinguishing simple from non-simple finite groups. At over 99% precision, the requisite answers can be estimated without recourse to standard computations which are many orders of magnitude slower.

It is therefore natural to wonder whether the elliptic curve data can be "machine-learned". Of course, we need to be careful. While computational algebraic geometry over $\mathbb{C}$ hinged on finding kernels and cokernels of integer matrices, a task in which AI excels. Problems in number theory are much less controlled. Indeed, trying to predict prime numbers [He] seems like a hopeless task, as mentioned in the introduction. Nevertheless, let us see how far we can go with our present dataset for BSD.

5.1 Predicting from the Weierstraß Coefficients

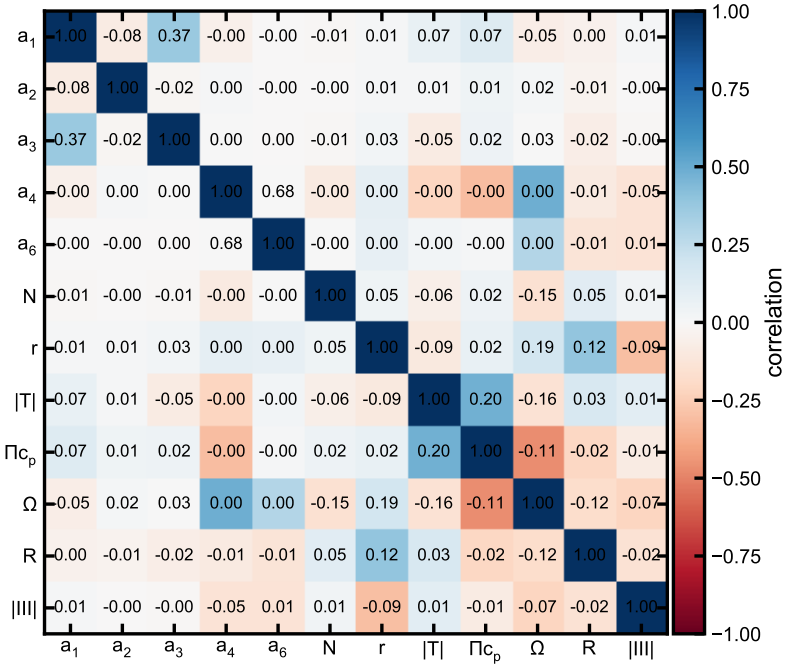


Figure 10: **Correlation matrix for the quantities characterizing ellipses.** Colours are assigned based on the value of the Pearson correlation coefficient between all pairs of quantities. The strength of the correlation is also reported for each pair.

We begin with quantifying the performance of machine learning models in predicting, one by one, the quantities N, $\prod_{p|N} c_p$, R, $|Ш|$, Ω, r, $|T|$, together with the RHS of the Conjecture 2, given the Weierstraß coefficients a_1, a_2, a_3, a_4, a_6 alone. Straight-away, this is expected to be a difficult, if not impossible task (as impossible as, perhaps, the prediction of prime numbers). This is confirmed by the correlation between the Weierstraßcoefficients and the BSD quantities: from the correlation matrix, we see that the relationship is indeed weak (cf. Fig. 10), implying this is not a

straightforward prediction task. Here, the correlation matrix contains the values of the Spearman correlation [Pear], that is the Pearson correlation [Pear] between the rank variables associated to each pairs of the original variables.

Quantity	**NMAE (XGBoost)**	**NMAE (Dummy)**	**NMAE (Linear)**
N	24.842 ± 0.032	25.175 ± 0.026	25.158 ± 0.027
$\prod_{p\|N} c_p$	0.028 ± 0.006	0.077 ± 0.016	0.058 ± 0.012
R	0.075 ± 0.015	0.112 ± 0.023	0.108 ± 0.022
$\|Ш\|$	0.023 ± 0.015	0.044 ± 0.028	0.043 ± 0.027
Ω	3.120 ± 0.099	6.057 ± 0.189	6.016 ± 0.189
RHS Conj. 2	7.070 ± 0.238	7.548 ± 0.255	7.533 ± 0.250
	RMSE (XGBoost)	**RMSE (Dummy)**	**RMSE (Linear)**
N	114687.179 ± 63.171	115784.768 ± 78.329	115774.283 ± 78.302
$\prod_{p\|N} c_p$	273.912 ± 18.665	286.522 ± 19.679	285.731 ± 19.711
R	13.579 ± 0.886	14.201 ± 0.552	14.197 ± 0.555
$\|Ш\|$	6.797 ± 1.550	6.369 ± 1.794	6.524 ± 1.688
Ω	0.449 ± 0.001	0.584 ± 0.001	0.583 ± 0.001
RHS Conj. 2	4.300 ± 0.002	4.554 ± 0.004	4.526 ± 0.003

Table 1: **Performance of the regression models.** The Normalized Median Absolute Error ($NMAE$) and the Root Mean Squared Error ($RMSE$), for XGboost (left column), the dummy regressor (central column) and a linear regression (right column). The reported values are averages across 5-fold cross-validations, with the corresponding standard deviations.

Our analysis relies on gradient boosted trees [BST], using the implementation of XGboost [XGBoost], an open-source scalable machine learning system for tree boosting used in a number of winning Kaggle solutions (17/29 in 2015). In Appendix B, we present the similar results using a support vector machine, another highly popular machine-learning model, and see that the XGboost indeed performs better. Furthermore, based on the learning curves of the XGboost models (discussed in Appendix A), we have chosen a $5-$**fold cross-validation**, such that the training set includes 80% of the values, and the validation set the remaining 20%.

5.1.1 Numerical Quantities

First, we train regression models to predict the values of N, $\prod_{p|N} c_p$, R, $|Ш|$ and Ω. We evaluate the performance of the regression, by computing the normalized median absolute error:

$$NMAE = \frac{median(|Y_i - \hat{Y}_i|)}{max(Y_i) - min(Y_i)} \ , \tag{5.10}$$

where Y_i are the observed values and $\hat{Y}_i$ are the predicted values, and the rooted mean squared error:

$$RMSE = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n}} \ , \tag{5.11}$$

where n is the size of the test set. We desire that both NMAE and RMSE to be close to 0 for a good prediction.

We compare the result of the XGBoost regression with two baselines: (1) a linear regression model and (2) a dummy regressor, that always predicts the mean of the training set. We find that, in all cases, the machine learning algorithms perform significantly better than the baseline models (see Table 1) with respect to the NMAE and RMSE. However, XGboost performs only marginally better than the baselines in predicting the value of N. We report also the so-called Imimportance of the features for the XGBoost regressor in Fig. 15. Here, importance indicates how useful each feature is in the construction of the boosted decision trees, and is calculated as the average importance across trees. For a single tree, the importance of a feature is computed as the relative increase in performance resulting from the tree splits based on that given feature [XGBoost].

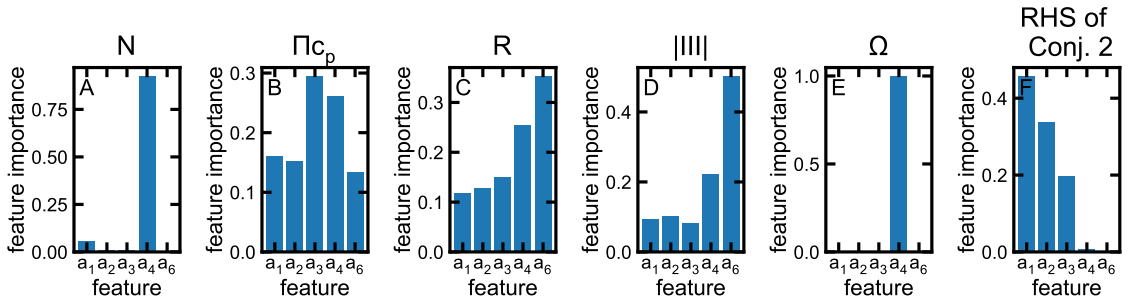


Figure 11: **Feature importance.** Feature importance of the XGBoost regression models for predicting N (A), $\prod_{p|N} c_p$ (B), R (C), $|Ш|$ (D) and Ω (E).

We see that overall our measures, NMAE and RMSE, are not too close to 0, except perhaps $|Ш|$, for which even a simple linear regression does fairly well. In table 2, we report the values of the coefficients of the linear regression fit for $|Ш|$. What Table 2

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5946	0.004	380.761	0.000	1.586	1.603
a_1	0.0658	0.005	14.524	0.000	0.057	0.075
a_2	-0.0065	0.004	-1.543	0.123	-0.015	0.002
a_3	-0.0518	0.005	-11.473	0.000	-0.061	-0.043
a_4	-0.6320	0.006	-110.282	0.000	-0.643	-0.621
a_6	0.4877	0.006	85.112	0.000	0.477	0.499

Table 2: **Prediction of $|Ш|$.** Coefficients of the linear regression for each of the features (inputs), with the associated standard deviation, the value of the t statistics and corresponding p-value. Here, we can reject the null hypothesis that the coefficients are equal to 0 at significance level $\alpha = 0.01$, for all coefficients, except the one associated to a_2 ($p > 0.01$).

means is that $|Ш| \simeq 1.5946 + 0.0658a_1 - 0.0065a_2 - 0.0518a_3 - 0.6320a_4 + 0.4877a_6$.

Predicted variable	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)
N	0	0	95.67	3.86e-101
$\prod_{p\|N} c_p$	0.006	0.006	2777	0
R	0.001	0.001	387.2	0
$\|Ш\|$	0.005	0.005	2522	0
Ω	0.005	0.005	2377	0
RHS Conj. 2	0.012	0.012	6143	0

Table 3: **Statistics of the linear regression models.** We report the R squared, the adjusted R squared, the F statistics, and the p-value associated to the F statistics for the different linear models. When the p-value of the F-statistics is close to 0, we can reject the null hypothesis that the intercept-only model provides a better fit than the linear model.

Likewise, in table 3, we report the the statistics associated to the linear regression models for the prediction of the various quantities, where only the Weierstrass coef-

ficients are used as features (inputs). The low R-squared values indicated that the regression is not so good in terms of the a_i coefficients alone.

5.1.2 Categorical Quantities

Next, we train classifiers to predict the values of r and $|T|$ because these easily fall into discrete categories: the rank $r = 0, 1, 2, 3$ and the torsion group size $|T|$ can only be one of 16 integer values due to Mazur's theorem. Again, we use a $5 - fold$ cross validation, and we evaluate the performance of the classifier, by computing the $F1$ score:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \; ; \quad \text{precision} := \frac{TP}{TP + FP} \; , \; \text{recall} := \frac{TP}{TP + FN} \tag{5.12}$$

where we have, in the predicted versus actual, the true positives (TP), false positives (FP), and false negatives (FN). Since we have several possible values for the rank r, we compute both $F1_{micro}$, by counting the total TP, FN and FP, as well as $F1_{macro}$, the average value of $F1$ computed across ranks.

In addition, we also compute the Matthew correlation coefficient MCC [Matthew], to describe the confusion matrix:

$$MCC := \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \; . \tag{5.13}$$

Both F1-scaore and MCC are desired to be close to 1 for a good prediction.

For checks, we compare the XGBoost classifier with (1) a baseline classifier, that predicts always the predominant class in the training set, as well as with (2) a logistic regression. We find that XGboost performs better than the baseline models for predicting $|T|$, but the performance of the prediction of r is comparable to the baselines (see Table 4).

Analyzing the confusion matrices (see Figures 12 and 13), it appears clear that it is very hard to predict $|T|$ and r from the Weierstraß coefficients alone. For both the prediction of r and $|T|$, the most important predictor is a_4 (Figures 12 and 13). This is the feature that contributed the most to increase the performance of the boosted tree [XGBoost].

Quantity	$F1_{micro}$ **(XGBoost)**	$F1_{micro}$ **(Dummy)**	$F1_{micro}$ **(Logistic)**
r	0.502 ± 0.001	0.502 ± 0.001	0.502 ± 0.001
$\|T\|$	0.582 ± 0.001	0.543 ± 0.001	0.518 ± 0.001
	$F1_{macro}$ **(XGBoost)**	$F1_{macro}$ **(Dummy)**	$F1_{macro}$ **(Logistic)**
r	0.179 ± 0.001	0.167 ± 0.001	0.167 ± 0.001
$\|T\|$	0.097 ± 0.001	0.059 ± 0.001	0.080 ± 0.001
	MCC **(XGBoost)**	MCC **(Dummy)**	MCC **(Logistic)**
r	0.0172 ± 0.0006	0.0000 ± 0.0000	-0.0002 ± 0.0010
$\|T\|$	0.1871 ± 0.0010	0.0000 ± 0.0000	0.0299 ± 0.0012

Table 4: **Performance of the classification models.** The scores $F1_{micro}$, $F1_{macro}$, and the Matthew correlation coefficient MCC, for XGboost (left column), the dummy regressor (central column) and a logistic regression (right column). The reported values are averages across 5-fold cross-validations, with the corresponding standard deviations.

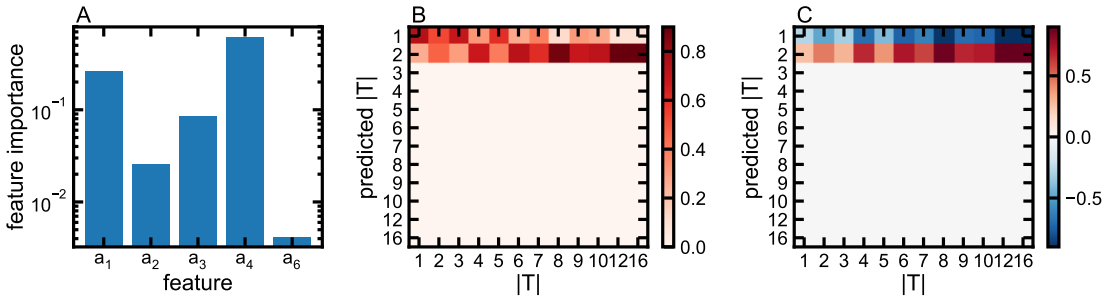


Figure 12: **Prediction of $|T|$.** (A) Importance of the different features (inputs) to predict $|T|$. (B) Confusion matrix (normalized by column) showing the fraction of entries $|T|$ with given *predicted* $|T|$. (C) Difference between the confusion matrix obtained for the XGBoost and the dummy classifier. Results are averaged over a 5-fold cross validation.

5.2 Mixed Predictions

While the results in the previous subsection may seem disappointing in general, they do present a good sanity check: to obtain all the BSD quantities from the elliptic curve data in some straight-forward way would be an almost unimaginable feat in number theory. Nevertheless, in this section, let us build machine learning models to predict

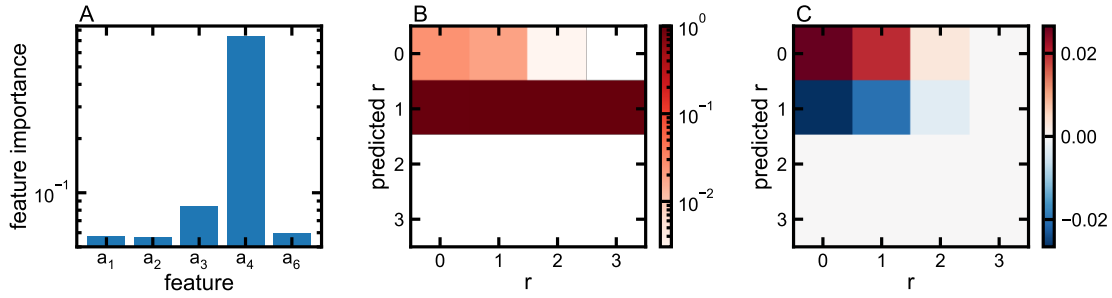


Figure 13: **Prediction of r.** (A) Importance of the different features to predict r. (B) Confusion matrix (normalized by column) showing the fraction of entries with rank r with given *predicted* r. (C) Difference between the confusion matrix obtained for the XGBoost and the dummy classifier. Results are averaged over a 5-fold cross validation.

the values of N, $\prod_{p|N} c_p$, R, $|Ш|$, Ω, r and $|T|$ among themselves, i.e., we consider as features (inputs) all the quantities characterizing the elliptic curves (except the predicted quantity), rather than the Weierstraß coefficients alone.

Quantity	**NMAE (XGBoost)**	**NMAE (Dummy)**	**NMAE (Linear)**
N	23.426 ± 0.031	25.175 ± 0.026	24.408 ± 0.039
$\prod_{p\|N} c_p$	0.012 ± 0.003	0.077 ± 0.016	0.065 ± 0.014
R	0.014 ± 0.003	0.112 ± 0.023	0.089 ± 0.018
$\|Ш\|$	0.006 ± 0.004	0.044 ± 0.028	0.048 ± 0.031
Ω	2.343 ± 0.103	6.057 ± 0.189	5.324 ± 0.174

Table 5: **Performance of the regression models considering as features all the quantities characterizing an ellipses.** The normalized median absolute error $NMAE$, for XGboost (left column), the dummy regressor (central column) and a linear regression (right column). The reported values are averages across 5-fold cross-validations, with the corresponding standard deviations. Results are considerably improved compared to table 1.

We present the results in Table 5 of the accuracy measure NMAE by the 3 methods as in subsection 5.1: machine-learning by XGBoost, dummy regression and linear regression. To read the table, of each of the 5 quantities given in a row, we use the other 4 to train the ML in order to predict this row. We see that this is a significant improvement over the previous subsection and shows that, especially the XGBoost, the machine-learning can very confidently predict the Tamagawa number,

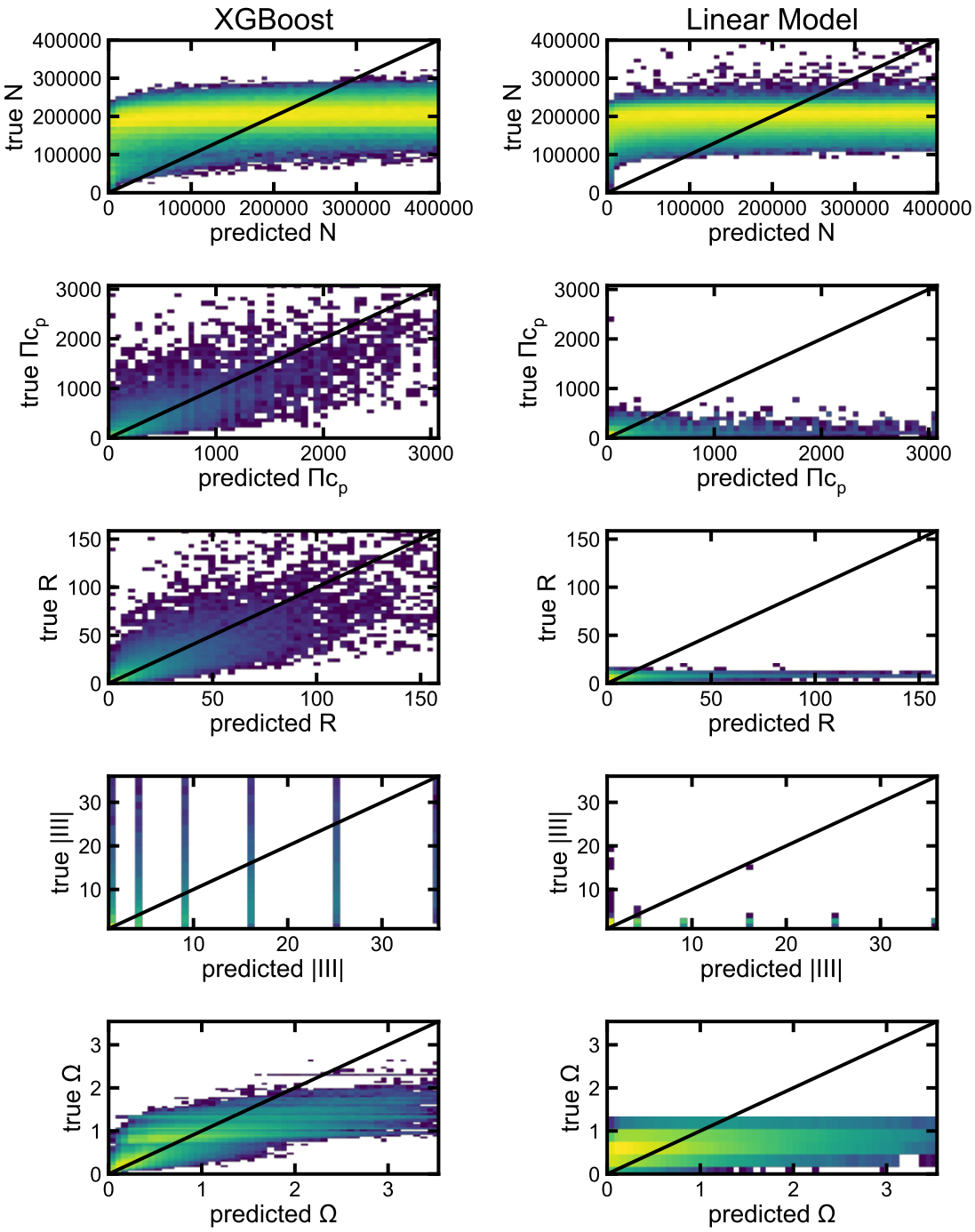


Figure 14: **True vs Predicted values** Results are shown for all the quantities, using XGBoost (left column) and the Linear model (right column).

the regulator and |Ш|. The feature importance is shown in Fig. 15 and in table 6, we report the the statistics associated to the mixed predictions linear regression models.

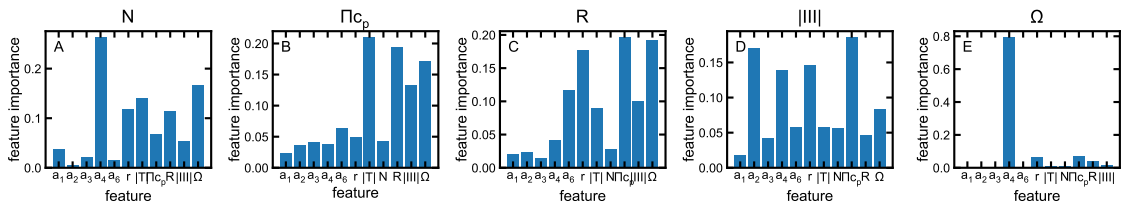


Figure 15: **Feature importance considering as features all the quantities characterizing an elliptic curve.** Feature importance of the XGBoost regression models for predicting N (Part A), $\prod_{p|N} c_p$ (Part B), R (Part C), $|Ш|$ (Part D) and Ω (Part E).

A comparison between the predictions of the linear models compared to XGboost is presented in Figure 14.

Predicted variable	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)
N	0.038	0.038	8999	0
$\prod_{p\|N} c_p$	0.054	0.054	12960	0
R	0.042	0.042	9889	0
$\|Ш\|$	0.017	0.017	3891	0
Ω	0.114	0.114	29110	0

Table 6: **Statistics of the linear regression models (mixed predictions).** We report the R squared, the adjusted R squared, the F statistics, and the p-value associated to the F statistics for the mixed predictions linear models.

Finally, let us use all quantities: the coefficients a_i as well as N, $\prod_{p|N} c_p$, R, $|Ш|$, Ω to predict r and $|T|$. The accuracy measure $F1$ and MCC (which should be close to 1 ideally) are presented in Table 7 and the feature importance, in Figures 16 and 17. We see that these are considerably improved compared to those obtained in section 5.1 (cf. Tables 1 and 4). This is somehow to be expected in light of the correlations observed in Figure 10. In fact, even logistic regressions behave fairly well, with the $F1_{micro}$, $F1_{macro}$, and the MCC scores subtiantially larger than those obtained by the dummy classifiers (see Table 7). For reference, we include report the hyperplane equations of the linear regression models to see how each of the quantities can be

Quantity	$F1_{micro}$ **(XGBoost)**	$F1_{micro}$ **(Dummy)**	$F1_{micro}$ **(Logistic)**
r	0.900 ± 0.001	0.502 ± 0.001	0.730 ± 0.001
$\lvert T\rvert$	0.908 ± 0.001	0.543 ± 0.001	0.567 ± 0.001
	$F1_{macro}$ **(XGBoost)**	$F1_{macro}$ **(Dummy)**	$F1_{macro}$ **(Logistic)**
r	0.554 ± 0.001	0.167 ± 0.001	0.387 ± 0.001
$\lvert T\rvert$	0.585 ± 0.015	0.059 ± 0.001	0.090 ± 0.001
	MCC **(XGBoost)**	MCC **(Dummy)**	MCC **(Logistic)**
r	0.8311 ± 0.0005	0.0000 ± 0.0000	0.5240 ± 0.0011
$\lvert T\rvert$	0.8302 ± 0.0014	0.0000 ± 0.0000	0.1364 ± 0.0009

Table 7: **Performance of the classification models considering as features all the quantities characterizing an elliptic curve.** The scores $F1_{macro}$, $F1_{micro}$ and the Matthew correlation coefficient MCC, for XGboost (left column), the dummy regressor (central column) and a logistic regression (right column). The reported values are averages across 5-fold cross-validations, with the corresponding standard deviations. Results are considerably improved compared to table 4.

fitted by all the others:

$$
\begin{aligned}
N &= 195500.0000 - 1526.8435a_1 + 288.1786a_2 - 806.9174a_3 - 122.4328a_4 + 16.3690a_6 + \\
&\quad 8047.4199r - 10250.0000|T| + 2192.0941 \prod_{p|N} c_p + 3197.1580R + 1293.3455|Ш| - 20330.0000\Omega \\
\prod_{p|N} c_p &= 49.7747 + 14.9069a_1 + 3.6860a_2 + 2.1424a_3 - 2.4121a_4 + 1.0613a_6 + \\
&\quad 17.8877r + 53.2012|T| + 5.3471N - 14.5038R - 4.3689|Ш| - 27.7937\Omega \\
R &= 4.0689 - 0.0028a_1 - 0.1430a_2 - 0.2379a_3 - 0.2995a_4 + 0.1230a_6 + \\
&\quad 2.1850r + 0.4585|T| + 0.3910N - 0.7271 \prod_{p|N} c_p - 0.2167|Ш| - 2.1082\Omega \\
|Ш| &= 1.5946 + 0.0517a_1 + 0.0094a_2 - 0.0195a_3 - 0.6322a_4 + 0.4875a_6 - \\
&\quad 0.5472r + 0.0112|T| + 0.0756N - 0.1046 \prod_{p|N} c_p - 0.1035R - 0.3466\Omega \\
\Omega &= 0.6065 - 0.0252a_1 + 0.0113a_2 + 0.0160a_3 - 0.0030a_4 + 0.0019a_6 + \\
&\quad 0.1147r - 0.0717|T| - 0.0945N - 0.0530 \prod_{p|N} c_p - 0.0801R - 0.0276|Ш|
\end{aligned}
\tag{5.14}
$$

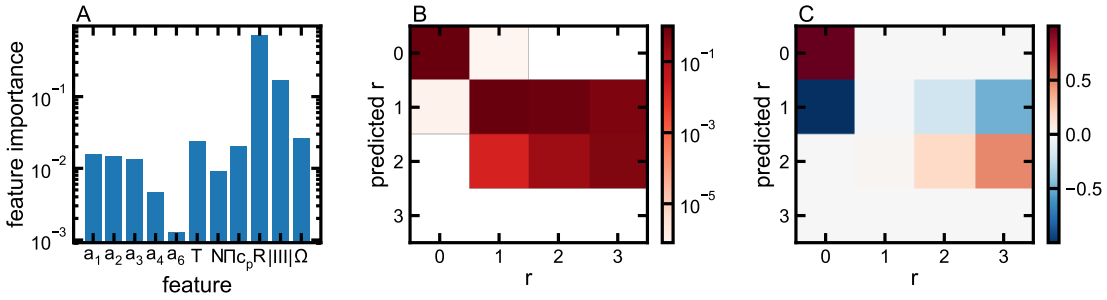


Figure 16: **Prediction of r considering as features all the quantities characterizing an ellipses.** (A) Importance of the different features to predict r. (B) Confusion matrix (normalized by column) showing the fraction of entries with rank r with given *predicted* r. (C) Difference between the confusion matrix obtained for the XGBoost and the dummy classifier. Results are averaged over a 5-fold cross validation.

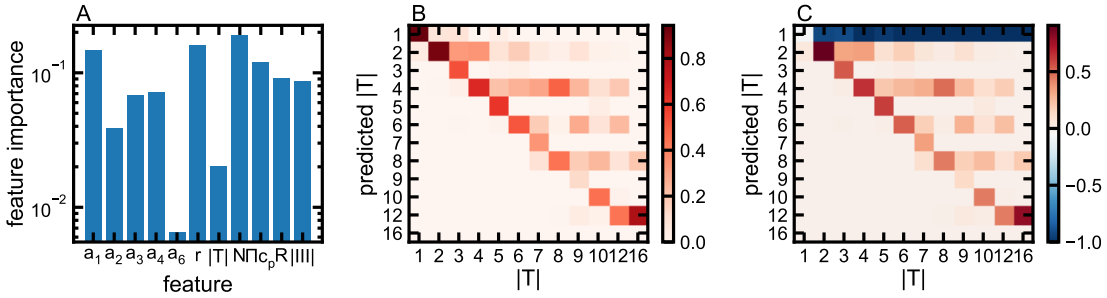


Figure 17: **Prediction of $|T|$ considering as features all the quantities characterizing an elliptic curve.** (A) Importance of the different features to predict $|T|$. (B) Confusion matrix (normalized by column) showing the fraction of entries with value $|T|$ and given *predicted* $|T|$. (C) Difference between the confusion matrix obtained for the XGBoost and the dummy classifier. Results are averaged over a 5-fold cross validation.

6 Conclusions and Prospects

In this paper, we initiated the study of the data science of the arithmetic of elliptic curves in light of the Birch-Swinnerton-Dyer Conjecture. This is inspired by the the recent advances in the statistical investigation of Calabi-Yau manifolds, especially in the context of super-string theory [HJP, ACHN], as well as in the paradigm of machine-learning structures in geometry [He, HeBook] and algebra [HK]. Whilst we are still within the landscape of "Calabi-Yau-ness", it is expected that patterns in number theory should be much more subtle than those in geometry over $\mathbb{C}$ and

in combinatorics. Nevertheless, BSD, residing at the crux between arithmetic and analysis, might be more susceptible to machine-learning and to pattern-recognition.

From our preliminary examinations on the extensive database of Cremona [Cre, Cre2] we have already found several interesting features. First, we find that in the minimal Weierstraß representation, where a pair of coefficients (a_4, a_6) clearly constitutes the principle component of the data, the distribution thereof follows a curious cross-like symmetry across rank, as shown in Figures 1 and 2. This is a highly-non-trivial symmetry since $a_{4,6} \leftrightarrow \pm a_{4,6}$ does not preserve rank. This symmetry is reminiscent of mirror symmetry for Calabi-Yau threefolds. In addition, the absence of data-points beyond the boundaries of the cross is also of note, much like that Hodge plot for the Calabi-Yau threefolds.

Over all, the distribution of the Euclidean distance of (a_4, a_6) to the origin, as well as that of the RHS of the Strong BSD, viz., the quantity $\frac{|Ш|\cdot\Omega\cdot R\cdot \prod_{p|N} c_p}{|T|^2}$ (cf. conjectures in Sec. 2.2), are best described by a Beta-distribution, which is selected in both cases among 85 continuous distributions using the Akaike Information Criterion. Organized by rank, these distributions also vary.

One further visualize the data, the tuples consisting of the coefficients $(a_1, a_2, a_3, a_4, a_6)$ as well as the BSD tuple $(N,\ |T|,\ \prod_{p|N} c_p,\ \Omega,\ R, Ш)$ for ranks $r = 0, 1, 2$ using the standard techniques from topological data analysis. The bar-codes are shown in Figures 6 and 7. While the Weierstraßcoefficients show little variation over rank, the BSD tuple does show differences over r. Moreover, as expected, the divisibility of the conductor N influences the barcodes.

Finally, emboldened by the recent success in using machine-learning to computing bundle cohomology on algebraic varieties without recourse to sequence-chasing and Gr$\acute{o}$bner bases as well as recognizing whether a finite group is simple directly by looking at the Cayley table, we asked the question of whether one can "predict" quantities otherwise difficult to compute directly from "looking" at the shape of the elliptic curve. Ideally, one would have hoped that training on a_i, one could predict any of the BSD quantities to high precision, as was in the cohomology case. However, due to the very high variation in the size of a_i, one could not find a good machine-learning technique, decision trees, support-vector machines or neural networks, which seems to achieve this. This is rather analogous to the (expected) failure of predicting

prime numbers using AI. Nevertheless, the BSD quantities, when *mixed* with the Weierstraß coefficients, does behave well under machine-learning. For instance, the Matthew correlation coefficient between predicted and true values of r and $|T|$ is ~ 0.83.

At some level, the experiments here, in conjunction with those in [He, KS, Rue, CHKN,BHJM,JKP,HK,BCDL,**?**], tend to show a certain *hierarchy of difficulty* in how machine-learning responds to problems in mathematics. Understandably, number theory is the most difficult: as a reprobate, [He] checked that trying to predict the next prime number, for instance, seems unfeasible for simple neural networks. On the other hand, algebraic geometry over the complex numbers seems to present a host of amenable questions such as bundle cohomology, recognition of elliptic fibrations or calculating Betti numbers. In between lie algebra and combinatorics, such as the structure of finite groups, where precision/confidence of the cross-validation is somewhat intermediate. It is therefore curious that in this paper one sees that a problem such as BSD, which resides in between arithmetic and geometry, is better behaved under machine-learning than a direct attack on patterns in primes.

Acknowledgments

YHH would like to thank the Science and Technology Facilities Council, UK, for grant ST/J00037X/1, Nankai University, China for a chair professorship and Merton College, Oxford, for enduring support.

Appendices

A Learning Curves

To prevent overfitting, we compute the learning curves of the regression (figure 18) and classification (figure 19) models in section 5.1. We find that 80% of the data is a good choice for the training set size, suggesting a 5-fold cross validation.

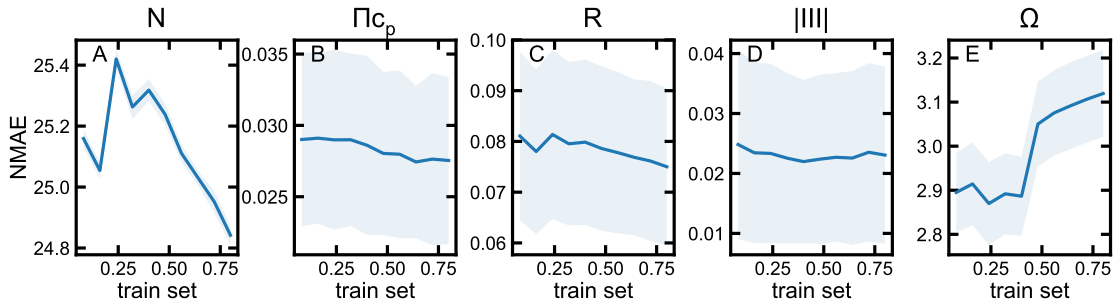


Figure 18: **Learning curves for the regression models.** The Normalized Median Absolute Error as a function of the train set size of the XGBoost regression models for predicting N (A), $\prod_{p|N} c_p$ (B), R (C), $|\text{Ш}|$ (D) and Ω (E). The shaded areas correspond to standard deviation across a $5-fold$ cross validation.

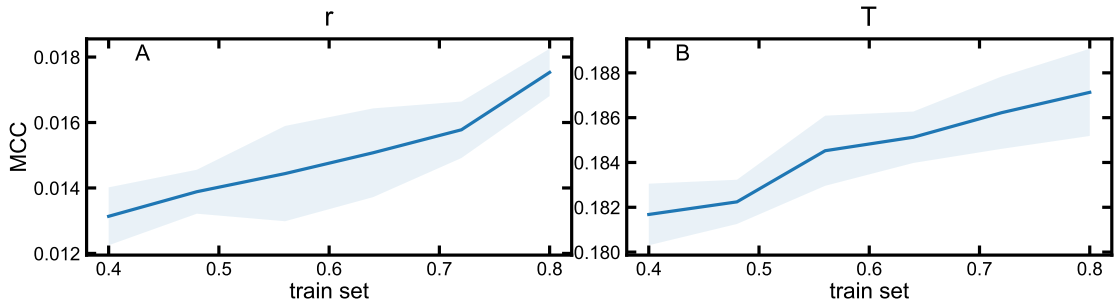


Figure 19: **Learning curves for the classification models.** The Matthew coefficient as a function of the train set size of the XGBoost regression models for predicting r (A) and $|T|$ (B). The shaded areas correspond to standard deviation across a $5-fold$ cross validation.

B Comparison with SVM

In this section, we compare the performance of the XGBoost models with Support Vector Machine (SVM) models. SVM models are very long to train, hence we focus, for this task, on a subset of $100,000$ examples. In table 8 we report the performance of the regression models used to predict N, $\prod_{p|N} c_p$, R, $|Ш|$ and Ω. Only in the case of Ω, the SVM model performs better than XGBoost.

Quantity	**NMAE (XGBoost)**	**NMAE (Dummy)**	**NMAE (SVM)**
N	114881.964 ± 364.409	115664.993 ± 384.936	115690.292 ± 417.181
$\prod_{p\|N} c_p$	278.372 ± 34.807	273.775 ± 27.528	275.182 ± 27.412
R	17.493 ± 4.178	15.124 ± 4.137	15.417 ± 4.067
$\|Ш\|$	4.938 ± 1.156	4.868 ± 1.223	4.893 ± 1.218
Ω	0.498 ± 0.004	0.584 ± 0.005	0.607 ± 0.005

Table 8: **Performance of the regression models.** The normalized median absolute error $NMAE$, for XGboost (left column), the dummy regressor (central column) and Support Vector Machine Regression (right column). The reported values are averages across 5-fold cross-validations, with the corresponding standard deviations.

C Characteristics of the Weierstraß coefficients.

a1	a2	a3	rank	size	$\overline{a_4}$	s_{a_4}	median	zero entries
0	-1	0	0	126135	-5E+09	7E+11	-8E+03	98
1	1	1	0	67834	-8E+10	7E+12	-2E+04	54
1	1	0	0	69759	-2E+11	3E+13	-1E+04	47
1	0	1	0	71309	-9E+10	1E+13	-2E+04	35
1	0	0	0	66411	-1E+12	2E+14	-2E+04	42
1	-1	1	0	96995	-1E+11	2E+13	-2E+04	41
0	1	1	0	18016	-4E+10	3E+12	-2E+03	38
0	1	0	0	118942	-1E+10	1E+12	-9E+03	108
0	0	1	0	28440	-1E+11	2E+13	-3E+03	546
1	-1	0	0	102769	-2E+11	5E+13	-2E+04	97
0	0	0	0	172238	-6E+09	9E+11	-1E+04	832
0	-1	1	0	17238	-1E+10	1E+12	-2E+03	40
0	-1	1	1	24593	-1E+09	1E+11	-1E+03	65
1	-1	0	1	127198	-1E+11	2E+13	-1E+04	150

1	0	0	1	98092	-5E+11	1E+14	-7E+03	77
0	1	1	1	27360	-5E+10	4E+12	-1E+03	54
1	0	1	1	94595	-3E+11	6E+13	-7E+03	62
1	-1	1	1	128957	-2E+10	2E+12	-9E+03	40
0	0	0	1	213780	-5E+10	2E+13	-6E+03	962
0	1	0	1	157003	-8E+09	1E+12	-5E+03	164
1	1	0	1	88403	-5E+10	4E+12	-7E+03	107
0	-1	0	1	155604	-1E+10	2E+12	-5E+03	159
0	0	1	1	39235	-5E+10	7E+12	-2E+03	608
1	1	1	1	91717	-3E+10	4E+12	-7E+03	87
1	1	0	2	18293	-2E+08	2E+10	-1E+03	28
1	0	0	2	25286	-5E+07	2E+09	-2E+03	23
1	-1	1	2	28940	-5E+07	6E+09	-2E+03	17
0	-1	0	2	30236	-3E+07	2E+09	-1E+03	44
1	0	1	2	20907	-2E+08	2E+10	-1E+03	17
0	0	0	2	40731	-6E+07	6E+09	-2E+03	126
1	-1	0	2	25793	-6E+08	8E+10	-2E+03	46
1	1	1	2	21197	-7E+07	5E+09	-1E+03	23
0	0	1	2	11187	-2E+07	9E+08	-5E+02	96
0	1	1	2	9609	-1E+08	1E+10	-5E+02	19
0	-1	1	2	7582	-7E+06	2E+08	-3E+02	22
0	1	0	2	34585	-2E+07	9E+08	-1E+03	36
1	-1	0	3	551	-8E+04	1E+06	-3E+02	1
0	0	0	3	698	-2E+04	2E+05	-3E+02	0
1	1	0	3	496	-2E+04	2E+05	-2E+02	0
0	0	1	3	506	-2E+04	2E+05	-2E+02	2
0	-1	1	3	399	-8E+04	1E+06	-2E+02	1
0	1	0	3	722	-8E+03	4E+04	-4E+02	1
0	-1	0	3	659	-1E+04	6E+04	-3E+02	0
1	0	0	3	612	-6E+03	3E+05	-4E+02	1
0	1	1	3	426	4E+03	9E+05	-3E+02	1
1	-1	1	3	604	-1E+04	7E+04	-4E+02	3
1	0	1	3	548	-1E+04	1E+05	-3E+02	3
1	1	1	3	458	-2E+04	4E+05	-2E+02	0
1	-1	0	4	1	-8E+01	NAN	-8E+01	0

Table 9: For given values of a_1,a_2,a_3, and r, the table reports the number of elliptic curves (size), and some statistics of the Weierstraß coefficient a_4 including the mean ($\overline{a_4}$), the standard deviation (s_{a_4}), the median (median) and the number of zero entries (zero entries)

a1	a2	a3	rank	size	$\overline{a_6}$	s_{a_6}	median	zero entries
0	-1	0	0	126135	-9E+15	4E+18	-5E+02	217
1	1	1	0	67834	-3E+17	9E+19	-1E+03	1
1	1	0	0	69759	-3E+18	6E+20	-7E+02	202

1	0	1	0	71309	2E+17	2E+20	-2E+03	2
1	0	0	0	66411	-4E+19	1E+22	-5E+03	176
1	-1	1	0	96995	-6E+17	3E+20	-3E+03	1
0	1	1	0	18016	-2E+17	3E+19	-4E+02	2
0	1	0	0	118942	-2E+16	6E+18	-1E+03	226
0	0	1	0	28440	-3E+18	5E+20	-6E+02	1
1	-1	0	0	102769	7E+18	2E+21	-8E+02	206
0	0	0	0	172238	1E+16	6E+18	-8E+02	486
0	-1	1	0	17238	-1E+16	4E+18	-2E+02	1
0	-1	1	1	24593	1E+15	2E+17	1E+01	17
1	-1	0	1	127198	-2E+18	6E+20	-4E+00	271
1	0	0	1	98092	-3E+19	9E+21	5E+01	246
0	1	1	1	27360	-1E+17	4E+19	-2E+00	18
1	0	1	1	94595	1E+19	3E+21	2E+01	27
1	-1	1	1	128957	-2E+14	1E+19	4E+01	15
0	0	0	1	213780	-1E+18	6E+20	-5E-01	604
0	1	0	1	157003	-3E+15	5E+18	2E+01	281
1	1	0	1	88403	-1E+17	4E+19	0E+00	271
0	-1	0	1	155604	2E+15	1E+19	0E+00	301
0	0	1	1	39235	6E+17	1E+20	-2E+01	24
1	1	1	1	91717	1E+17	4E+19	7E+00	17
1	1	0	2	18293	8E+13	1E+16	1E+03	42
1	0	0	2	25286	4E+12	4E+14	1E+04	42
1	-1	1	2	28940	-1E+13	3E+15	1E+04	21
0	-1	0	2	30236	7E+11	3E+14	2E+03	58
1	0	1	2	20907	7E+13	9E+15	3E+03	22
0	0	0	2	40731	-2E+13	3E+15	4E+03	82
1	-1	0	2	25793	7E+14	1E+17	2E+03	57
1	1	1	2	21197	-9E+12	1E+15	5E+03	23
0	0	1	2	11187	1E+12	1E+14	1E+03	26
0	1	1	2	9609	-4E+13	4E+15	2E+03	18
0	-1	1	2	7582	4E+10	8E+12	6E+02	24
0	1	0	2	34585	5E+11	9E+13	5E+03	42
1	-1	0	3	551	1E+08	3E+09	2E+03	0
0	0	0	3	698	-2E+06	2E+08	2E+03	0
1	1	0	3	496	3E+06	9E+07	9E+02	1
0	0	1	3	506	-6E+06	2E+08	1E+03	4
0	-1	1	3	399	1E+08	3E+09	6E+02	3
0	1	0	3	722	1E+06	1E+07	3E+03	0
0	-1	0	3	659	3E+06	3E+07	2E+03	0
1	0	0	3	612	-5E+06	2E+08	3E+03	1
0	1	1	3	426	1E+08	2E+09	1E+03	4
1	-1	1	3	604	3E+06	4E+07	3E+03	4
1	0	1	3	548	5E+06	6E+07	2E+03	4
1	1	1	3	458	5E+07	1E+09	1E+03	2
1	-1	0	4	1	3E+02	NAN	3E+02	0

Table 10: For given values of a_1,a_2,a_3, and r, the table reports the number of curves (size), and some statistics of the Weierstraß coefficient a_6 including the mean ($\overline{a_6}$), the standard deviation (s_{a_6}), the median (median) and the number of zero entries (zero entries)

References

[ACHN] R. Altman, J. Carifio, J. Halverson and B. D. Nelson, "Estimating Calabi-Yau Hypersurface and Triangulation Counts with Equation Learners," JHEP **1903**, 186 (2019) doi:10.1007/JHEP03(2019)186 [arXiv:1811.06490 [hep-th]].

[BCDL] C. R. Brodie, A. Constantin, R. Deen and A. Lukas, "Machine Learning Line Bundle Cohomology," arXiv:1906.08730 [hep-th].

[BHJM] K. Bull, Y. H. He, V. Jejjala and C. Mishra, "Machine Learning CICY Threefolds," Phys. Lett. B **785**, 65 (2018) [arXiv:1806.03121 [hep-th]].

–, "Getting CICY High," Phys. Lett. B **795**, 700 (2019) [arXiv:1903.03113 [hep-th]].

[BS] M. Bhargava, C. Skinner, "A positive proportion of elliptic curves over Q have rank one", arXiv:1401.0233.

[BSD] B. Birch, P. Swinnerton-Dyer, "Notes on Elliptic Curves (II)". J. Reine Angew. Math. 165 (218): 79 - 108 (1965).

[BST] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[BSZ] M. Bhargava, C. Skinner,W. Zhang, "A majority of elliptic curves over Q satisfy the Birch and Swinnerton-Dyer conjecture", arXiv:1407.1826

[BGZ] J. Buhler, B. Gross, D. Zagier, "On the Conjecture of Birch and Swinnerton-Dyer for an Elliptic Curve of Rank 3", Mathematics of Computation, Vol. 44, No. 170, 1985, pp. 473 - s481

[Ca] Calabi, Eugenio, "The space of Kähler metrics", Proc. Internat. Congress Math. Amsterdam, 2, pp. 206 - 207 (1954)

– "On Kähler manifolds with vanishing canonical class", in Fox, Spencer, Tucker, *Algebraic geometry and topology. A symposium in honor of S. Lefschetz*, Princeton Mathematical Series, 12, PUP, pp. 78 - 89 (1957).

[CHKN] J. Carifio, J. Halverson, D. Krioukov and B. D. Nelson, "Machine Learning in the String Landscape," JHEP **1709**, 157 (2017) doi:10.1007/JHEP09(2017)157 [arXiv:1707.00655 [hep-th]].

[Cre] John Cremona, "The Elliptic Curve Database for Conductors to 130000", in: Hess F., Pauli S., Pohst M. (eds) *Algorithmic Number Theory*. ANTS 2006. Lecture Notes in Computer Science, vol 4076. Springer.

[Cre2] John Cremona, "The L-functions and modular forms database project",

arXiv:1511.04289, http://www.lmfdb.org/

[CZCG] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, "Persistence barcodes for shapes," Intl. J. Shape Modeling, 11 (2005), 149-187.

[Doug] M. R. Douglas, "The Statistics of string / M theory vacua," JHEP **0305**, 046 (2003) [hep-th/0303194].

[Ei] Greg Helsenman, Eirene, http://gregoryhenselman.org/eirene/

[Elk] N. Elkies, "Three lectures on elliptic surfaces and curves of high rank", Lecture notes, Oberwolfach, 2007, arXiv:0709.2908

[He] Y. H. He, "Deep-Learning the Landscape," arXiv:1706.02714 [hep-th].
Y. H. He, "Machine-learning the string landscape," Phys. Lett. B **774**, 564 (2017).

[HeBook] Y. H. He, "The Calabi-Yau Landscape: from Geometry, to Physics, to Machine-Learning," arXiv:1812.02893 [hep-th].

[HJP] Y. H. He, V. Jejjala and L. Pontiggia, "Patterns in CalabiâĂŞYau Distributions," Commun. Math. Phys. **354**, no. 2, 477 (2017) [arXiv:1512.01579 [hep-th]].

[HK] Y. H. He and M. Kim, "Learning Algebraic Structures: Preliminary Investigations," arXiv:1905.02263 [cs.LG].

[HL] Y. H. He and S. J. Lee, "Distinguishing Elliptic Fibrations with AI," arXiv:1904.08530 [hep-th].

[JKP] V. Jejjala, A. Kar and O. Parrikar, "Deep Learning the Hyperbolic Volume of a Knot," arXiv:1902.05547 [hep-th].

[KS] D. Krefl and R. K. Seong, "Machine Learning of Calabi-Yau Volumes," Phys. Rev. D **96**, no. 6, 066014 (2017) [arXiv:1706.03346 [hep-th]].

[Las] M. Laska, "An Algorithm for Finding a Minimal Weierstrass Equation for an Elliptic Curve", Maths. of computation, V. 38, no. 157, 1982.

[Maz] B. Mazur, "Modular curves and the Eisenstein ideal". Publications Mathématiques de l'IHÉS. 47 (1): 33 - 186, 1977

[Mil] J. Milne, "Elliptic curves,", http://www.jmilne.org/math/Books/ectext5.pdf

[OPTGH] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod and Heather A Harrington, "A roadmap for the computation of persistent homology" EPJ Data Science20176:17, https://arxiv.org/abs/1506.08903

[Pear] Sheskin, D. J., Handbook of parametric and nonparametric statistical procedures. Chapman and Hall/CRC (2003).

[PJ] Peres-Neto, P. R., Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia, 129(2), 169-178.

[Rue] F. Ruehle, "Evolving neural networks with genetic algorithms to study the String Landscape," JHEP **1708**, 038 (2017) [arXiv:1706.07024 [hep-th]].

[RS] K. Rubin , A. Silverberg "Ranks of elliptic curves", Bull. AMS 39 (2002), 455 - 474.

[Sil] J. Silverman, "The arithmetic of elliptic curves", GTM 106, Springer-Verlag (1992).

[Tat] J. T. Tate, "The arithmetic of elliptic curves," Invent. Math, 23, 1974, pp 179-206

[TW] R. Taylor, A. Wiles, "Ring-theoretic properties of certain Hecke algebras", Ann. of Math. (2) 141 (1995), no. 3, 553 - 572.

[XGBoost] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[Yau] S.-T. Yau, "Calabi's conjecture and some new results in algebraic geometry," Proc. Nat. Acad., USA, 74 (5), pp 1798-9, (1977)

–, "On the Ricci curvature of a compact KÃďhler manifold and the complex Monge-AmpÃÍre equation I", Comm. Pure and Applied Maths, 31 (3), pp 339-411, (1978).

[Matthew] Gorodkin, Jan. "Comparing two K-category assignments by a K-category correlation coefficient." Computational biology and chemistry 28.5-6 (2004): 367-374.